Published online 3 May 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W695–W699***
doi:10.1093/nar/gkq313

A new bioinformatics analysis tools framework at EMBL–EBI

Mickael Goujon, Hamish McWilliam, Weizhong Li, Franck Valentin, Silvano Squizzato, Juri Paern and Rodrigo Lopez*

European Bioinformatics Institute, EMBL Outstation, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received January 27, 2010; Revised April 6, 2010; Accepted April 17, 2010

ABSTRACT

The EMBL-EBI provides access to various mainstream sequence analysis applications. These include sequence similarity search services such as BLAST, FASTA, InterProScan and multiple sequence alignment tools such as ClustalW, T-Coffee and MUSCLE. Through the sequence similarity search services, the users can search mainstream sequence databases such as EMBL-Bank and UniProt, and more than 2000 completed genomes and proteomes. We present here a new framework aimed at both novice as well as expert users that exposes novel methods of obtaining annotations and visualizing sequence analysis results through one uniform and consistent interface. These services are available over the web and via Web Services interfaces for users who require systematic access or want to interface with customized pipe-lines and workflows using common programming languages. The framework features novel result visualizations and integration of domain and functional predictions for protein database searches. It is available at http://www.ebi.ac.uk/Tools/sss for sequence similarity searches and at http://www.ebi.ac.uk/Tools/msa for multiple sequence alignments.

INTRODUCTION

Bioinformatics is a vast and complex multidisciplinary research area where numerous tools have been developed over the years to analyse constantly growing amounts of data. Since 1998, the European Bioinformatics Institute (EMBL–EBI) has provided public access to various mainstream sequence analysis applications (1,2). These include sequence similarity search services (http://www.ebi.ac.uk/Tools/similarity.html), such as FASTA (3), BLAST (4,5) and InterProScan (6) and multiple sequence alignment tools (http://www.ebi.ac.uk/Tools/sequence.html), such as ClustalW (7), T-Coffee (8), MUSCLE (9), Kalign (10) and MAFFT (11). These services are provided via a PERL-CGI job dispatcher framework for managing job submission and result representation. This infrastructure handled more than 16 million jobs during 2009. The popularity of these services has made it necessary to redesign the system in order to minimize maintenance and enhance the integration of features requested by users. A new and modular framework, called JDispatcher, has been developed to improve the accessibility and quality of the services relevant to the biological community.

JDispatcher framework

JDispatcher is aimed at both novice and expert users and exposes novel methods of obtaining annotations and visualizing sequence analysis results through one uniform and consistent interface. These services are available interactively over the web and via SOAP and REST interfaces for systematic or programmatic use. The new framework provides input validation to assure successful job submissions, offers new visualization features to assist in the interpretation of results and uses the EBI search engine, EB-eye (12), to integrate relevant annotations.

A user can submit sequences using web forms that contain all supported parameters and their possible values. The different tools have been grouped into categories based on their purpose (Table 1).

Within a category, the tools share the same interface design, which uses well established usability patterns, such as wizard-like steps to guide the user through the submission process. It makes use of decision-trees to validate all the parameters required to warrant successful job submissions. If the validation fails, the user is notified about which specific parameters or data are invalid, and the job is not submitted. Alternatively, JDispatcher

*To whom correspondence should be addressed. Tel: +44 1223 494423; Fax: +44 1223 494468; Email: rls@ebi.ac.uk

assigns a unique job identifier and sends a request to a workload management system for the job to be executed. The identifier is then used to keep track of the tasks and to retrieve the results when they become available. The results of each job are kept for a maximum of 7 days.

Results representation

The results of an analysis are made available using various representations (e.g. HTML tables, XML files, images, etc.). In order to produce these representations, each result is converted into a generic category-specific model that is used by a renderer that generates the requested output. The renderers are specific to the model and not to the tool, and thus are available across all the tools in a category. The availability of multiple views of the same data helps the user to interpret and compare results from different tools within a category.

Sequence search algorithms produce limited hits annotation. With the new framework it is possible to navigate hits and access related information. Figure 1 shows the 'Summary Table' of an SSEARCH of mouse glomulin (UniProtKB/Swiss-Prot GLMN_MOUSE), which is essential for the development of the vascular system, against the UniProtKB/Swiss-Prot database (13). Each column heading has clickable arrows that allow the user to sort the results according to the values in the columns [e.g. sequence length, score, percentage identity, positives and *E*()-value]. Each match is enriched with links to cross-references and related information in various data resources (e.g. gene expression, genomic sequences, structures, function, ontologies and literature citations). Optionally, the alignment from the search, and/or the full-annotation for the selected matches can be displayed. A hits selection can also be downloaded in fasta format.

Table 1. Tools available in the JDispatcher framework

Category	Tool
Sequence Similarity Searches (sss)	psisearch, psiblast, ncbiblast, wublast, fasta, ssearch, ggsearch and glsearch
Multiple Sequence Alignments (msa)	clustalw2, tcoffee, kalign, muscle, mafft, and prank

Align.	DB:ID	Source	Length	Score	Identities	Positives	E()
1	SP:GLMN_MOUSE	Glomulin OS=Mus musculus GN=Glmn PE=2 SV=1 *Cross-references and related information in:* ▶ Gene Expression ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature ▶ Protein Sequences	596	3881	100.0	100.0	0.0
2	SP:GLMN_HUMAN	Glomulin OS=Homo sapiens GN=GLMN PE=1 SV=2 *Cross-references and related information in:* ▶ Gene Expression ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Molecular Interactions ▶ Protein Families ▶ Literature ▶ Protein Sequences	594	3307	85.6	95.0	0.0
3	SP:ALF4_ARATH	Aberrant root formation protein 4 OS=Arabidopsis thaliana GN=ALF4 PE=2 SV=2 *Cross-references and related information in:* ▶ Gene Expression ▶ Nucleotide Sequences ▶ Ontologies ▶ Protein Families ▶ Literature ▶ Protein Sequences	626	168	21.5	52.6	0.0039
4	SP:PSLS_METJA	Phosphosulfolactate synthase OS=Methanocaldococcus jannaschii GN=comA PE=1 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Ontologies ▶ Enzymes ▶ Protein Families ▶ Literature ▶ Macromolecular Structures	251	136	25.6	59.4	0.22
5	SP:MON2_XENLA	Protein MON2 homolog OS=Xenopus laevis GN=mon2 PE=2 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Ontologies ▶ Protein Families	1721	148	22.5	51.7	0.36
6	SP:MON2_HUMAN	Protein MON2 homolog OS=Homo sapiens GN=MON2 PE=1 SV=2 *Cross-references and related information in:* ▶ Gene Expression ▶ Nucleotide Sequences ▶ Genomes ▶ Ontologies ▶ Protein Families ▶ Literature ▶ Protein Sequences	1718	142	25.6	56.5	0.95
7	SP:NDC80_SCHPO	Kinetochore protein ndc80 OS=Schizosaccharomyces pombe GN=ndc80 PE=1 SV=1 *Cross-references and related information in:* ▶ Gene Expression ▶ Nucleotide Sequences ▶ Ontologies ▶ Molecular Interactions ▶ Protein Families ▶ Literature	624	132	20.2	53.7	1.3
8	SP:TRMB_CLOB6	tRNA (guanine-N(7)-)-methyltransferase OS=Clostridium botulinum (strain 657 / Type Ba4) GN=trmB PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Ontologies ▶ Enzymes ▶ Protein Families	217	120	24.9	52.8	2.4
9	SP:TRMB_CLOBL	tRNA (guanine-N(7)-)-methyltransferase OS=Clostridium botulinum (strain Langeland / NCTC 10281 / Type F) GN=trmB PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Ontologies ▶ Enzymes ▶ Protein Families	217	118	24.9	52.8	3.3
10	SP:TRMB_CLOB1	tRNA (guanine-N(7)-)-methyltransferase OS=Clostridium botulinum (strain ATCC 19397 / Type A) GN=trmB PE=3 SV=1 *Cross-references and related information in:* ▶ Nucleotide Sequences ▶ Ontologies ▶ Enzymes ▶ Protein Families ▶ Literature	217	118	24.9	52.8	3.3

Figure 1. Summary Table view of the results obtained when searching the sequence of mouse glomulin against the UniProtKB/Swiss-Prot database using SSEARCH.

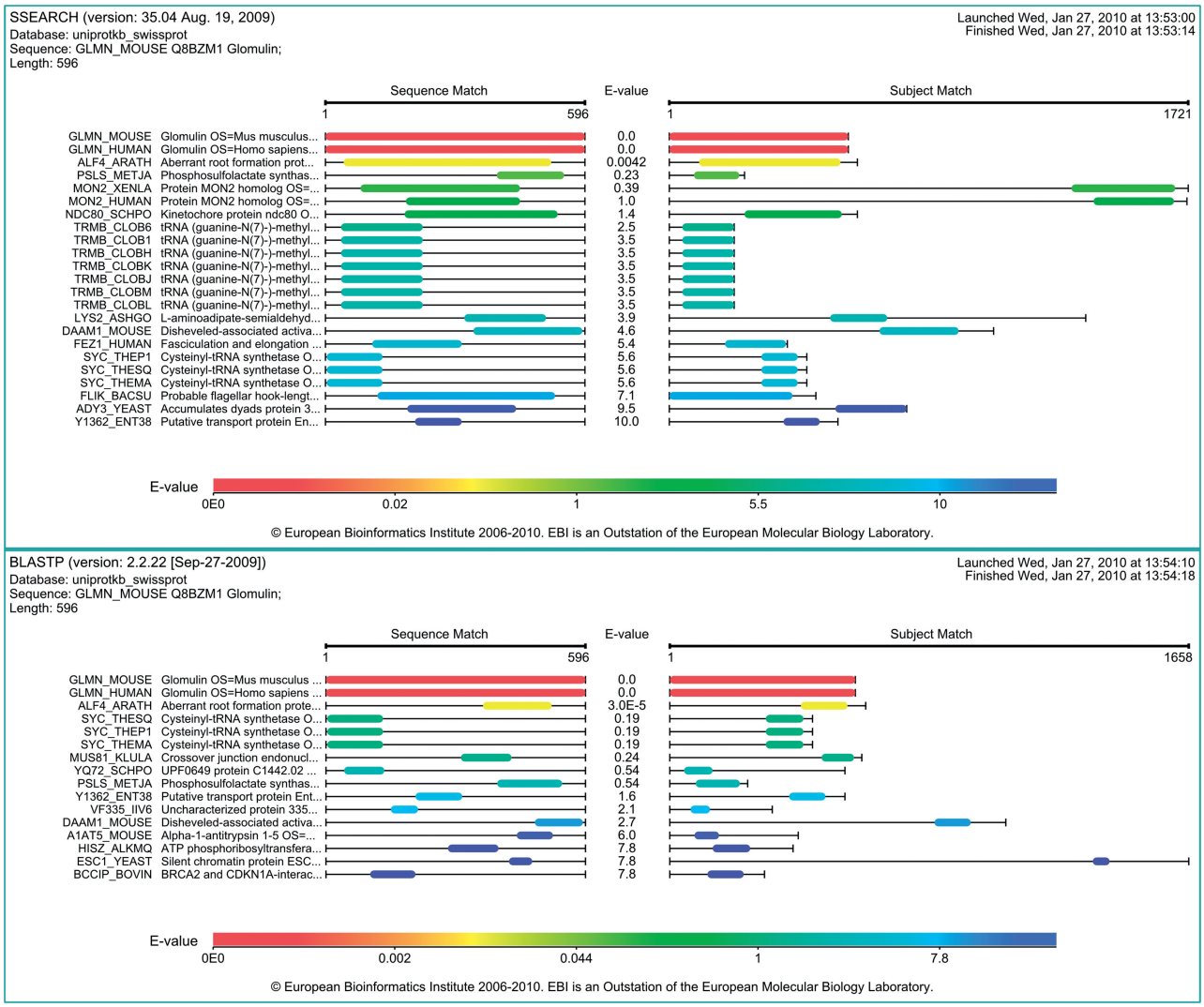

Figure 2. Comparisons between the Visual output results obtained when searching the sequence of mouse glomulin against the UniProtKB/Swiss-Prot database using SSEARCH and NCBI BLAST, respectively.

Figure 2 shows the 'Visual Output' obtained from searches using SSEARCH and NCBI BLAST of the glomulin sequence against UniProtKB/Swiss-Prot using default parameters. Comparison of the two images reveals notable differences in the sequence matches reported by the two search methods. For example, differences in the aligned regions between glomulin and aberrant root formation protein 4 for Arabidopsis (ALF4_ARATH) are clearly visible in both; SSEARCH identifies two MON2 homologues at *E*()-values <1 (MON2_XENLA and MON2_HUMAN), which may indicate there is a structural relationship between GLMN at the C-terminus of the MON2 homologues, although these may not share related functions.

Determining which functional domains and families a protein belongs to is critical to the understanding of the biological processes it may be involved in. This is important for the characterization of existing drug targets as well as in the identification of novel ones. Family and domain functional predictions have been built into the framework, using pre-calculated matches from the InterPro Consortium (14) data. This enables users, not only to search for sequence similarities when using the UniProt databases, but also to characterize the sequence query in terms of domain architectures that may elicit its function. Figure 3 shows 'Functional Predictions' for a hypothetical bioactive lysophospholipid that was compared against UniProtKB/Swiss-Prot using NCBI BLAST. The hypothetical sequence has several good homologues, all belonging to the GPCR rhodopsin-like superfamily, which are clearly seen. This indicates the query protein could represent a potential target for receptor-binding studies.

In both, the 'Visual Output' and 'Functional Predictions' result representations, the matches are coloured, from red to blue, according to *E*()-value, using a relative scale, from the most to the least significant hits within the result. An absolute scale, which ranges from *E*() = 0 to *E*() = 1.0, is also available. These aim to aid the user in deciding whether weak similarities may be biologically significant. These images are available in Scalable

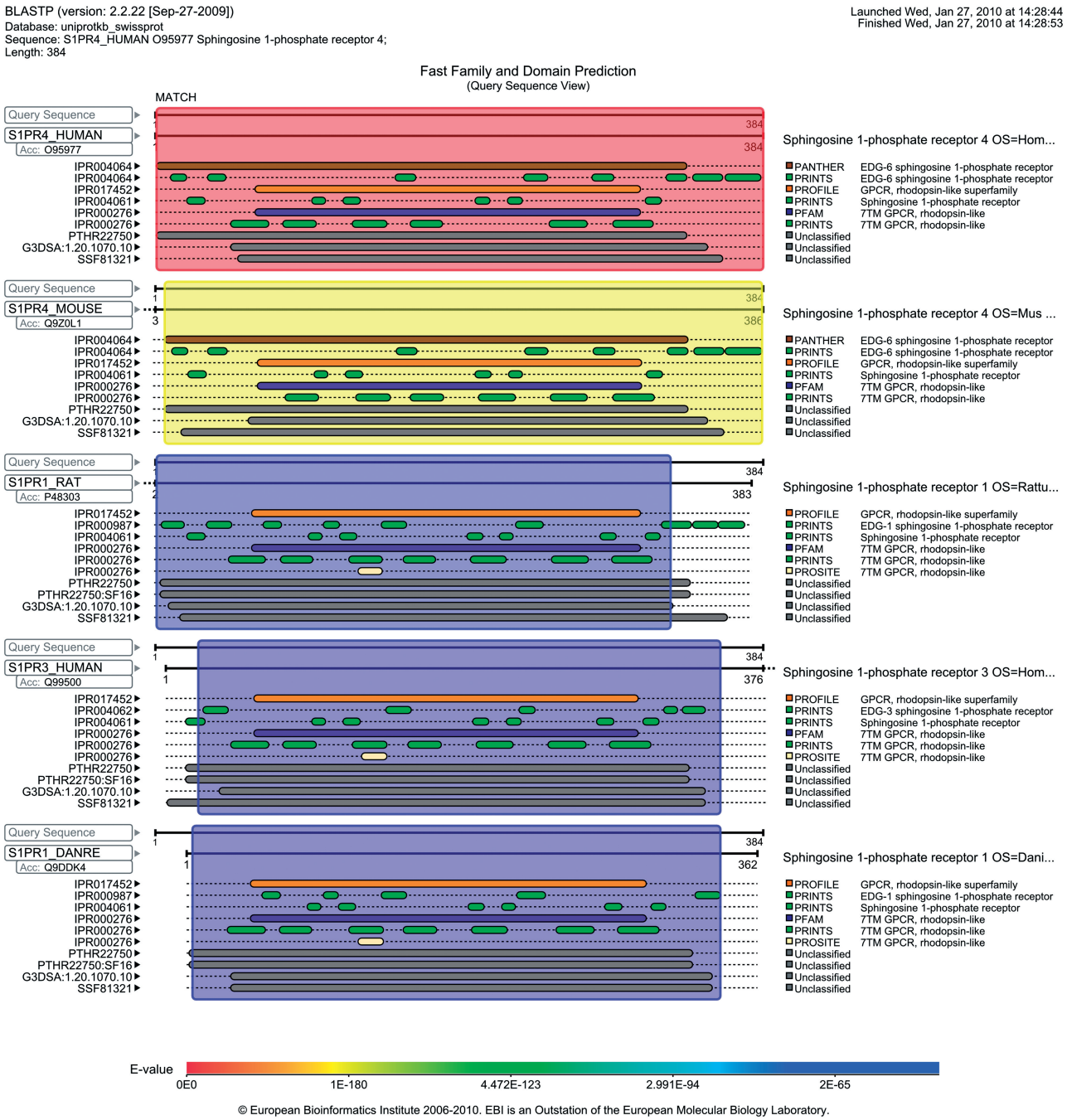

Figure 3. Functional prediction view of the results obtained when comparing the sequence of putative bioactive lysophospholipid that was compared against UniProtKB/Swiss-Prot using NCBI BLAST.

Vector Graphics (SVG), Portable Network Graphic (PNG) and JPEG output, providing wide compatibility. The raw result and processed forms, such as the 'Summary Table' content and XML formats are downloadable for further processing by the user.

The examples above illustrate how, from a single sequence similarity search, it is possible to access related sources of annotation, determine visually which results are relevant and infer gene and protein functional associations, using the JDispatcher framework.

Web Services

Web Services technologies have opened up important opportunities for the analysis of life sciences data. It is now well established that sharing resources, across geographically distributed networks, is advantageous to scientists and bioinformaticians through the re-use of generic services, such as those presented in this article. The new JDispatcher framework provides multiple front-ends: in addition to the web interface, SOAP and REST APIs (http://www.ebi.ac.uk/Tools/webservices/) have been

implemented to offer programmatic access using accepted web services standards.

The SOAP and REST APIs cater for users requiring systematic access to a wide range of sequence similarity search and multiple sequence alignment services, which can be built into local analytical workflows and pipelines (e.g. Taverna (15), Triana (http://www.trianacode.org/), KNIME (www.knime.org) (16) and Pipeline Pilot (http://accelrys.com/products/scitegic/index.html))—typical usage scenarios include the characterization of novel genomes and proteomes and the analysis of data derived from meta-genome experiments.

Using the APIs, complex applications can be developed in various programming languages, which include: C/C++, C#, Java, Perl, PHP, Python and Ruby, or scripting environments such a Bash, csh, batch and PowerShell. This allows integration of services into existing and/or new applications that require access to fast sequence database searching or multiple sequence alignment methods. To facilitate this type of usage, the services provide extensive meta-information describing the available parameters, including their possible values and descriptions of their purpose.

Typical applications of the JDispatcher framework services include: providing an alternative interface for specialist usage targeted at a specific community; integrating a service into an existing data portal to provide analysis services; and enhancing analysis results by directly connecting the result with the data. These are of importance to service providers and users of pipelines who may not have the resources to run and maintain the infrastructure required to support equivalent functionality.

CONCLUSIONS

The modularity of this new framework reduces maintenance overheads and simplifies the addition of tools and features. Keeping the result data model and the renderers separate provides the flexibility to add additional representations to all functionally related tools. This improves the level of usability for both novice and expert users. The presented visualization examples highlight important insights in the understanding of existing and new nucleotide and protein sequences from both genomes and metagenome experiments and suggest novel ways in which these data can be interpreted.

Academic and commercial laboratories can integrate the JDispatcher framework services with their local analytical pipelines or workflows. These represent an important contribution to the growing number of available services in bioinformatics and have been submitted to the BioCatalogue (17) (www.biocatalogue.org), a registry of freely available web services in the life sciences.

ACKNOWLEDGEMENTS

We acknowledge valuable feedback from Prof. William Pearson from the University of Virginia, USA and the InterPro and UniProt teams at EMBL-EBI.

FUNDING

The European Commission under FELICS [contract number 021902 (RII3), within the Research Infrastructure Action of the FP6 'Structuring the European Research Area' Programme]; core funding from the European Molecular Biology Laboratory; European Patent Office. Funding for open access charge: EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. McWilliam,H., Valentin,F., Goujon,M., Li,W., Narayanasamy,M., Martin,J., Miyar,T. and Lopez,R. (2009) Web services at the European Bioinformatics Institute—2009. *Nucleic Acids Res.*, **37**, W6–W10.
2. Brooksbank,C., Cameron,G. and Thornton,J. (2010) The European Bioinformatics Institute's data resources. *Nucleic Acids Res.*, **38**, D17–D25.
3. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
4. Lopez,R., Silventoinen,V., Robinson,S., Kibria,A. and Gish,W. (2003) WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res.*, **31**, 3795–3798.
5. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
6. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
7. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) ClustalW2 and ClustalX version 2.0. *Bioinformatics*, **23**, 2947–2948.
8. Notredame,C., Higgins,D. and Heringa,J. (2000) T-Coffee: a novel method for multiple sequence alignments. *J. Mol. Biol.*, **302**, 205–217.
9. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
10. Lassmann,T. and Sonnhammer,E.L. (2005) Kalign – an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, **6**, 298.
11. Katoh,K., Asimenos,G. and Toh,H. (2009) Multiple alignment of DNA sequences with MAFFT. *Methods Mol. Biol.*, **537**, 39–64.
12. Valentin,F., Squizzato,S., Goujon,M., McWilliam,H., Paern,J. and Lopez,R. (2010) Fast and efficient searching of biological data resources—using EB-eye. *Brief. Bioinformatics*, doi:10.1098/bib/bbp065 [Epub ahead of print 11 February 2010].
13. The UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.
14. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
15. Hull,D., Wolstencroft,K., Stevens,R., Goble,C., Pocock,M.R., Li,P. and Oinn,T. (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, **34**, W729–W732.
16. Berthold,M.R., Cebron,N., Dill,F., Gabriel,T.R., Kotter,T., Mein,T., Ohl,P., Sieb,C., Thiel,K. and Wiswedel,B. (2007) KNIME: The Konstanz Information Miner. *Data Analysis, Machine Learning and Applications – Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, Berlin, Germany, pp. 319–326.
17. Goble,C., Belhajjame,K., Tanoh,F., Bhagat,J., Wolstencroft,K., Stevens,R., Nzuobontane,E., McWilliam,H., Laurent,T. and Lopez,R. (2009) BioCatalogue: a curated web service registry for the life science community. *Nature Precedings*, http://www.iscb.org/uploaded/css/36/11627.pdf.